REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	15 Feb 00	Final Technical: 3/96-8/99

4. TITLE AND SUBTITLE
AI approaches to statistical language models

5. FUNDING NUMBERS
N00014-96-1-0549

6. AUTHOR(S)
Eugene Charniak

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brown University
Providence, RI 02912

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ONR
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Publicly available

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

We have made a number of advances under this grant. We have created what is currently the most accurate parser for parsing into Penn-tree-bank-style trees; a program that identifies the antecedents of pronouns with 85% accuracy; a program that assigns function tags to parse text with 85% accuracy; a program that assigns referents to full noun phrases with 65% accuracy; very efficient parsers -- parsers that explore very few constituents that do not appear in the final parse; and programs that discover semantic information about words from unlabeled text. Furthermore, all these programs work by statistical means.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Artificial intelligence, natural language processing, statistical NLP.	4
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std Z39-18
298-102

Final Report of ONR Grant N00014-96-1-0549

Eugene Charniak
Department of Computer Science
Brown University

February 15, 2000

1 Administrative Information

Title:	**AI Approaches to Statistical Language Models**
ONR Grant Number:	N00014-96-1-0549
Organization:	Brown University
Effective Date:	3/1/96
Closing Date:	8/31/99
Principal Investigator:	Eugene Charniak Professor of Computer Science (401) 863-7636 ec@cs.brown.edu

2 Summary of Results

In the proposal that resulted in this research grant we said, "We propose to create better statistical language models by bringing together statistical methodology and traditional AI approaches." It is fair to say that we have been successful in this goal. In this period we have created:

- what is currently the most accurate parser for parsing into Penn-treebank style trees. This parser has a per-constituent precision and recall of 89.5% (91.1% for sentences of length less than or equal to 40 words and punctuation) [Charniak 1999]
- a program that identifies the antecedents of pronouns with a 85% accuracy[Ge, Hale, Charniak 1998]

- a program that assigns function-tags to parse text with 85% accuracy. (E.g., the noun phrase "yesterday" would be given the function tag "temporal" in "He ate yesterday" to distinguish it from the role played by the noun phrase "pizza" in "He ate pizza") [Blaheta, Charniak 2000]

- a program that assigns referents to full noun phrases with a 65% accuracy [Hall, Charniak 2000]

- very efficient parsers — parsers that explore very few constituents that do not appear in the final parse [Caraballo, Charniak 1998] [Charniak, Goldwater, Johnson 1998] [Blaheta, Charniak 1999]

- programs that discover semantic information about words from unlabeled text [Roark, Charniak 1998] [Caraballo 1999] [Caraballo Charniak 1999] [Berland Charniak 1999]

Furthermore, as stated in our research proposal, all of these programs work by statistical means.

We have also been able to combine many of these programs in order to parse, find noun-phrase coreference, and function-tag, large quantities of text. For example, in the last month we have delivered to the LDC (Linguistic Data Consortium, the major organization for the distribution of large text and speech corpora) 35 million words of Wall Street Journal newspaper articles that have been machine annotated in the aforementioned fashion. The LDC will be distributing this new corpus.

We expect that in the years that follow we will be able to increase the accuracy and depth of this annotation. In particular we expect in the next few months to be able to provide not just a parse tree, but the predicate-argument structure of the sentences. More generally we hope to provide deeper semantic annotations such as case (or thematic) roles.

3 Publications

Parsing with context-free grammars and word statistics, Eugene Charniak, Technical Report CS-95-28, Department of Computer Science, Brown University (1995).

A statistical syntactic disambiguation program and what it learns, Eugene Charniak and Murat Ersan), TR CS-95-29 Brown University, Department of

Computer Science (1995). (Also appears in *Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing*, S. Wermter, E. Riloff, and G. Scheler Eds., New York: Springer pp 146-160 (1996).)

Figures of merit for best-first probabilistic chart parsing, Sharon Caraballo and Eugene Charniak, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp 127-132, (1996).

Taggers for parsers, Eugene Charniak, Glenn Carroll, John Adcock, Antony Cassandra, Yoshihiko Gotoh, Jeremy Katz, Michael Littman, and John McCann, *Artificial Intelligence*, V 28 No 1-2, pp 45-57 (1996).

Tree-bank grammars, Eugene Charniak, *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press: Menlo Park, pp. 1031-1036, (1996).

Expected-Frequency Interpolation Eugene Charniak, Technical Report CS96-37 Department of Computer Science, Brown University (1996)

Statistical parsing with a context-free grammar and word statistics Eugene Charniak, Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press, Menlo Park (1997)

Statistical Techniques for Natural Language Parsing Eugene Charniak, *AI Magazine*, Vol 18, No. 4, pp. 33-43 (1997)

New Figures of merit for best-first probabilistic chart parsing, Sharon Caraballo and Eugene Charniak, *Computational Linguistics* Vol. 4, pp. 275-298 (1998).

Edge-based best-first chart parsing, Eugene Charniak, Sharon Goldwater, and Mark Johnson, *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 127-133 (1998)

Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction, Brian Roark, and Eugene Charniak, pp. 1110-1116, *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, (1998)

A statistical approach to anaphora resolution, Niyu Ge, John Hale, and Eugene Charniak, *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 161-171, (1998)

Finding parts in very large corpora Matthew Berland, Eugene Charniak, *Proceedings of the ACL 1999*, pp. 57-64 (1999)

Automatic compensation for parser figure-of-merit flaws Don Blaheta, Eugene Charniak, *Proceedings of the ACL 1999*, pp. 513-518 (1999)

A maximum-entropy-inspired parser Eugene Charniak Technical Report CS99-12, Department of Computer Science, Brown University (1999) (Also accepted in Proceedings of the North-American Chapter of the Association for Computational Linguistics - NAACL-2000)

Determining the specificity of nouns from text Sharon Caraballo, Eugene Charniak, *1999 Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70 (1999)

Automatic construction of a hypernym-labeled noun hierarchy from text Sharon A. Caraballo, (1999)

A statistical method for noun-phrase coreference Keith Hall, Eugene Charniak, (2000)

Assigning function tags to text Don Blaheta, Eugene Charniak, *The 2000 Conference of the North American Chapter of the Association for Computational Linguistics* (2000)